

## CS523 - Big Data Technology – Final Project

### Part 1 :

**Context:** Covid19 data is ingested into system by Kafka, Spark Streaming read data from Kafka then store it in HBASE.

#### Prerequisites:

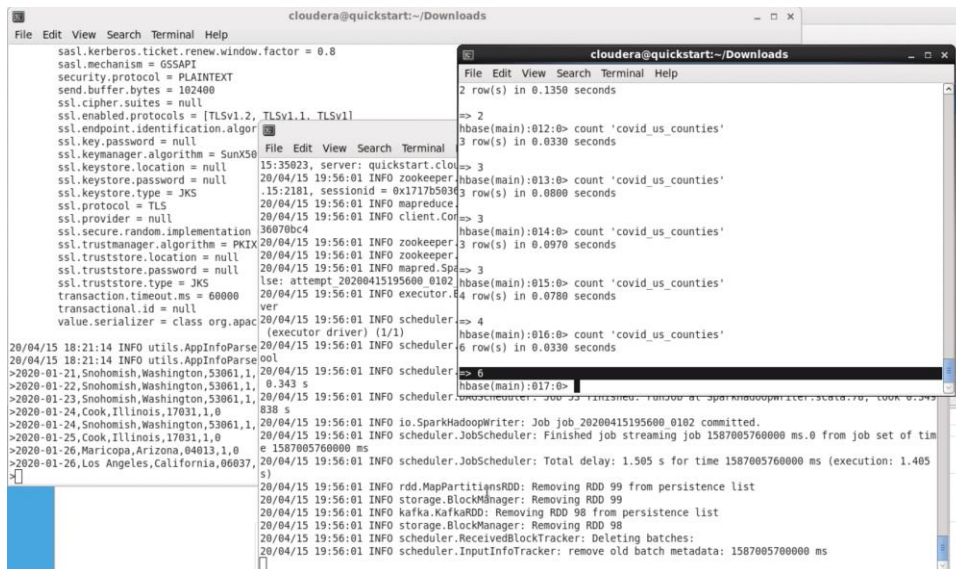
- Install Kafka as Part4
- Upgrade to Spark 2

**Java Class:** CovidDataFeed.java

**Running Command:** Running command: spark2-submit --master yarn --deploy-mode client --class CovidDataFeed finalproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar

**Shellscript:** covid\_streaming.sh

#### Screen shots:



The image shows two overlapping terminal windows from a Cloudera environment. The left window displays Spark logs for a job, including configuration details like 'ssl.enabled.protocols = [TLSv1.2, TLSv1.1, TLSv1]' and 'ssl.protocol = TLS'. It also shows a list of COVID-19 cases by state, such as '2020-01-21, Snohomish, Washington, 53061, 1'. The right window shows HBase console output where a user is querying the 'covid\_us\_counties' table. The output shows rows of data being returned, such as 'hbase(main):012:0> count 'covid\_us\_counties'' and '3 row(s) in 0.0330 seconds'.

### Part 2:

**Context:** Covid19 data has been downloaded from <https://github.com/nytimes/covid-19-data> and imported to "covid19\_us\_counties" table on HBase. We will use Spark SQL to perform query data on HBase table.

Import data to Hbase: hadoop jar finalproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar util.CovidHbaseTable

**Java file:** CovidSparkSQL.java

**Running command:** spark2-submit --master yarn --deploy-mode client --class CovidSparkSQL finalproject-0.0.1-SNAPSHOT-jar-with-dependencies.jar

**ShellScript:** covid\_streaming\_sql.sh

**Screen shots:**

```
20/04/14 09:41:24 INFO scheduler.DAGScheduler: Job 4 finished: show at Covid19Sql.java:29,
+-----+
| county|sum(cases)|
+-----+
| Snohomish|      5|
| Orange|      1|
| Los Angeles|      1|
| Cook|      2|
| Maricopa|      1|
+-----+
```

### Part 3:

**Context:** Covid19 data has been imported to “covid19\_us\_counties” table. We need to represent the data by visualize number of Covid19 cases in each county.

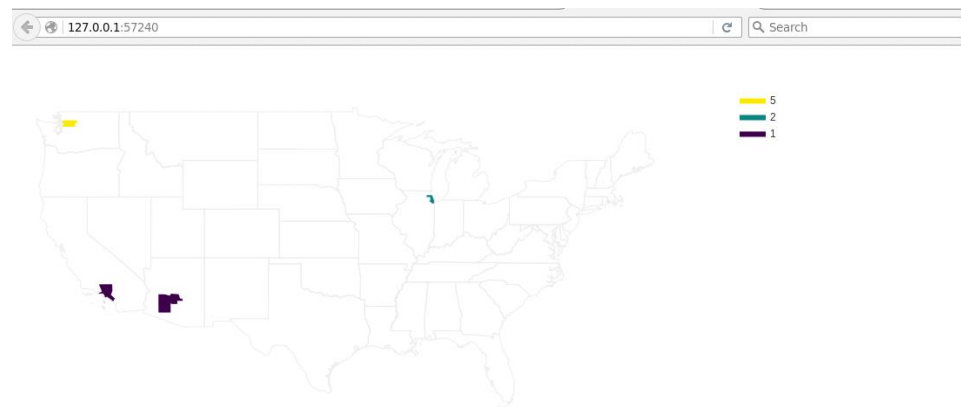
**Prerequisites:**

- Upgrade cloudera to support python 3
- Install happybase to connect with HBase table
- Install plotly and plotly-geo to visualize the data

**Python file:** covidvisual.py

**ShellScripts:** covidvisualdata.sh

**Screen shots:**



### Part 4: Kaka Install and Integrate with other services

**Install from parcels**

Parcel Name	Version	Status	Actions
ACCUMULO	1.7.2-5.5.0.ACCUMULO5.5.0.p0.8	Available Remotely	<a href="#">Download</a>
	1.4.4-1.cdh4.5.0.p0.65	Available Remotely	<a href="#">Download</a>
CDH 5	5.13.0-1.cdh5.13.0.p0.29	Distributed, Activated	<a href="#">Deactivate</a>
KAFKA	4.1.0-1.4.1.0.p0.4	Available Remotely	<a href="#">Download</a>
	3.1.1-1.3.1.1.p0.2	Distributed, Activated	<a href="#">Deactivate</a>

cloudera MANAGER

Clusters ▾

Hosts ▾

Diagnostics ▾

Audits ▾

Charts ▾

Administration ▾

📦

👤

Search

Support ▾

cloudera ▾

✓ Kafka-4 (Cloudera QuickStart)

Actions ▾

Apr 12, 8:43 AM PDT

Status

Instances

Configuration

Commands

Charts Library

Audits

Quick Links ▾

Filters

▼ STATUS

None 1

Good Health 1

► COMMISSION STATE

► MAINTENANCE MODE

► BACK

Search

Actions for Selected ▾

Add Role Instances

Role Groups

<input type="checkbox"/>	Role Type	State	Host	Commission State	Role Group
<input type="checkbox"/>	Gateway	N/A	<a href="#">quickstart.cloud<span>era</span></a>	Commissioned	Gateway Default Group
<input type="checkbox"/>	Kafka Broker (Active Controller)	Started	<a href="#">quickstart.cloud<span>era</span></a>	Commissioned	Kafka Broker Default Group

- kafka\_topic\_create.sh
- kafka\_producer.sh

```
cloudera@quickstart:~  
File Edit View Search Terminal Help  
ot be established. Broker may not be available.  
20/04/12 08:43:06 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:07 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:08 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:09 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:10 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:11 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:12 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:13 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:14 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:15 WARN clients.NetworkClient: [Producer clientId=console-producer] Connection to node 50 could not be established. Broker may not be available.  
Hi Kafka  
my test message  
cloudera@quickstart:~  
File Edit View Search Terminal Help  
20/04/12 08:43:11 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:12 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:13 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:14 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:15 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Connection to node 50 could not be established. Broker may not be available.  
20/04/12 08:43:16 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Error while fetching metadata with correlation id 5759 : (vtest)=INVALID REPLICATION FACTOR  
20/04/12 08:43:17 WARN clients.NetworkClient: [Consumer clientId=consumer-1, groupId=console-consumer-22584] Error while fetching metadata with correlation id 5760 : (vtest)=LEADER NOT AVAILABLE  
Hi Kafka  
my test message
```

## Install and configuration kafka

<https://blog.clairvoyantsoft.com/installing-apache-kafka-on-clouderas-quickstart-vm-8245d8d0ebe5>

Upgrade to spark 2

<https://blog.clairvoyantsoft.com/installing-spark2-on-clouderas-quickstart-vm-bbf0db5fb3a9>

Using Twitter api

<https://towardsdatascience.com/how-to-capture-and-store-tweets-in-real-time-with-apache-spark-and-apache-kafka-e5ccd17afb32>

<https://www.edureka.co/blog/spark-streaming/>

Databricks

<https://community.cloud.databricks.com/?o=688775375094051#setting/clusters>

Maven dependencies

[https://docs.cloudera.com/documentation/enterprise/release-notes/topics/cdh\\_vd\\_cdh5\\_maven\\_repo\\_57x.html#concept\\_970\\_mcn\\_yk](https://docs.cloudera.com/documentation/enterprise/release-notes/topics/cdh_vd_cdh5_maven_repo_57x.html#concept_970_mcn_yk)

Install Python 3

<https://stackoverflow.com/questions/45803713/safely-have-two-versions-of-python-on-cloudera-virtual-machine-without-python-in/45804331#45804331>

<https://stackoverflow.com/questions/35246386/conda-command-not-found>

```
conda create -y -n myproject 'python>3.6'
```

```
source activate myproject
```

```
python3
```

